Evaluating the complexity of some families of functional data

E. G. Bongiorno[1], A. Goia[2] and P. Vieu[3]

Abstract

In this paper we study the complexity of a functional data set drawn from particular processes by means of a two-step approach. The first step considers a new graphical tool for assessing to which family the data belong: the main aim is to detect whether a sample comes from a monomial or an exponential family. This first tool is based on a nonparametric kNN estimation of small ball probability. Once the family is specified, the second step consists in evaluating the extent of complexity by estimating some specific indexes related to the assigned family. It turns out that the developed methodology is fully free from assumptions on model, distribution as well as dominating measure. Computational issues are carried out by means of simulations and finally the method is applied to analyse some financial real curves dataset.

MSC: 62-09, 62G05, 60G99.

Keywords: Small ball probability, log-Volugram, random processes, complexity class, complexity index, knn estimation, functional data analysis.

1. Introduction

The description and the analysis of a statistical sample $X_1, \dots, X_n$ often rely on the complexity of the objects being observed. In usual multivariate situations (that is when each X_i is a d-dimensional vector) the degree of complexity is linked with the dimension d of the data which is in general known and statistical procedures are therefore developed to estimate and/or describe some probabilistic characteristic of the underlying random vector X (density function being the most common). For many reasons that we will discuss just below, this general approach cannot be followed in functional data analysis, that is the branch of statistics dealing with observations X_i which are curves, surfaces, images or other objects. Such a topic has attracted a lot of researchers and the interest towards this discipline is certified by monographs (see e.g. Bosq, 2000; Ferraty and

[1] DiSEI, Università del Piemonte Orientale, Via Perrone, 18, 28100, Novara, Italy. enea.bongiorno@uniupo.it

[2] DiSEI, Università del Piemonte Orientale, Via Perrone, 18, 28100, Novara, Italy. aldo.goia@uniupo.it

[3] Institut de Mathématiques de Toulouse, Université Paul Sabatier, France. philippe.vieu@math.univ-toulouse.fr

Received: June 2017
Accepted: February 2018

Vieu, 2006; Horváth and Kokoszka, 2012; Ramsay and Silverman, 2005), collections of recent contributions (see e.g. Aneiros et al., 2017; Bongiorno et al., 2014), special issues (see e.g. Kokoszka et al., 2017; Goia and Vieu, 2016) and recent articles (see among many others Bongiorno and Goia, 2016; Cardot, Cénac and Godichon-Baggioni, 2017; Chen, Delicado and Müller, 2017; Vilar, Raña and Aneiros, 2016). The question of defining the complexity of a functional sample has to be thought in a much more different way. The problem goes back to mathematical analysis in abstract infinite dimensional spaces, and more precisely to the difficulty for choosing some dominating measure (as could be the Lebesgue measure for continuous vectors or the counting measure for discrete ones). This is discussed in details for instance in Bogachev (1998). This has at least two important consequences. Firstly, the notion of density function has to be revisited, and secondly the notion of complexity of the model could not be reduced to a simple dimensionality index (see Bongiorno and Goia, 2017; Delaigle and Hall, 2010; Ferraty, Kudraszow and Vieu, 2012).

An usual way to overpass this difficulty when the sample comes from a variable X valued in some infinite dimensional topological space $\mathscr{F}$ is to consider the Small Ball Probability (SmBP), that is the asymptotic behaviour of $\mathbb{P}(X \in B(\chi,h))$ as h tends to zero. Here $B(\chi,h)$ stands for the ball centered at χ with radius h. Operatively, it is useful to assume that the SmBP satisfies for small h

$$\mathbb{P}(X \in B(\chi,h)) \sim \psi(\chi)\,\phi(h), \tag{1}$$

where, to ensure identifiability of the decomposition, one has to impose some normalization restriction like $\mathbb{E}[\psi(X)] = 1$. This factorization isolates the manner in which the SmBP depends upon χ and h through the *spatial* and *volumetric* terms ψ and ϕ respectively without referring to some dominating measure, and this justifies its utilization in literature (see for instance Gasser, Hall and Presnell, 1998 and Masry, 2005). Although the volumetric term has been studied extensively from a probabilistic point of view and mostly for the Gaussian processes (see the surveys on small tail literature Li and Shao, 2001; Lifshits, 2012 and references therein), from a statistical point it has only been used as a tool for controlling asymptotic behaviour of nonparametric functional estimator (see Ferraty and Vieu, 2006, Chapter 13; Masry, 2005). In fact, functional data analysis literature has focused mostly on the spatial term $\psi(\chi)$ since it naturally leads to define a surrogate density for the process and the methods vary from semi– to non–parametric approaches (see Bongiorno and Goia, 2016, 2017; Ciollaro et al., 2014; Delaigle and Hall, 2010; Delsol and Louchet, 2014; Ferraty et al., 2012) with applications in various statistical problems like defining/estimating functional modes (see Ferraty and Vieu, 2006, Chapter 6; Delaigle and Hall, 2010; Gasser et al., 1998) and classification problems (see Bongiorno and Goia, 2016; Ciollaro et al., 2014; Jacques and Preda, 2014).

To understand how the volumetric term ϕ can be of help in evaluating the complexity, firstly consider the multivariate setting $\mathscr{F} = \mathbb{R}^d$. Here, the complexity parameter is the dimension d which appears in $\phi(h) = v_d h^d$ (with v_d being the volume of the d–dimensional unit ball), while the function $\psi(\chi)$ represents the d–dimensional density function. In the functional setting, there is an important additional problem coming from the fact that the concentration function $\phi(\cdot)$ may be of many different forms, most of them being not as simple as in the multivariate one: often ϕ can not be expressed in closed form not even asymptotically (see Bongiorno and Goia, 2017; Delaigle and Hall, 2010). On the other hand, there are some remarkable cases whose volumetric term can be explicitly written; in particular, let us look at three specific cases (trajectories drawn from two of them are depicted in Figure 1):

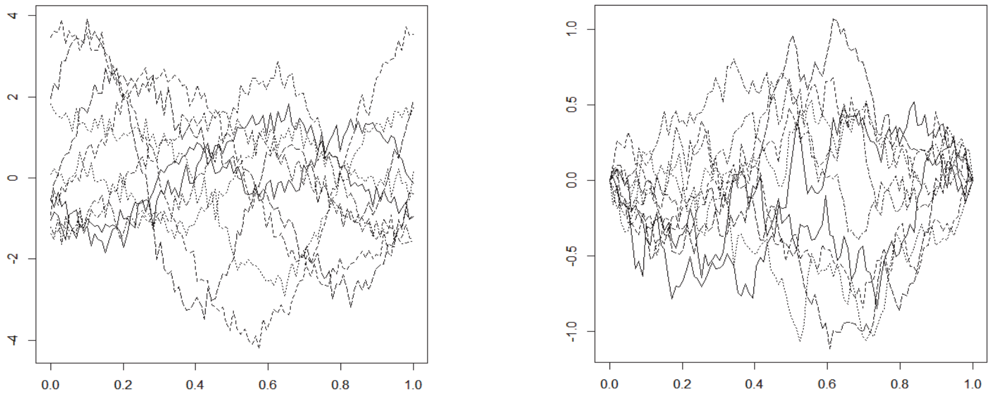

Figure 1: *Ten trajectories drawn from a noised 3–dimensional process and a Brownian Bridge process are depicted on left and right panel respectively.*

Case 1 The functional data have some finite dimensional structure. In this case the concentration function has the monomial form $\phi(h) = c_d h^d$, for some constant term c_d and the complexity of the model is the positive integer parameter d. This happens for instance when the topology on the functional space is constructed by looking only at d directions (of a given orthonormal basis) of the functional elements (see Ferraty and Vieu, 2006, Chapter 13).

Case 2 The functional data have some fractal structure (see Ferraty and Vieu, 2006, Definition 13.1). This is an extension of the first situation in which the concentration function takes the form $\phi(h) = c_\alpha h^\alpha$, for some constant term c_α and the complexity of the model is now the (non integer) positive parameter α.

Case 3 The functional data come from some Gaussian processes. This corresponds for instance to Wiener, Brownian Bridge or diffusion processes in $\mathscr{L}^2_{[0,1]}$ (see Li and Shao, 2001), for which the concentration function has the exponential form $\phi(h) = C_1 h^\gamma \exp\left\{-C_2/h^\beta\right\}$ with $\beta \in (0,\infty)$ and $\gamma \in [0,\infty)$. In this case the complexity of the data is measured by the indexes γ, β which cannot be interpreted as some dimensionality parameters (see Li and Shao, 2001 for deeper discussion and more examples of exponential type processes). Note that, such an exponential structure is implicitly linked with the existence of some Gaussian dominating measure for the process.

In what follows, the term *monomial* (*exponential* resp.) *family* refers to the set of processes like in Case 1 or 2 (Case 3, resp.). Since these cases are, to the best of our knowledge, the only ones for which the volumetric term can be specified and cover a wide range of situations, we limit our analysis to them.

It is worth noticing that, by analogy with the finite dimensional setting, the function ϕ may reveal some latent features of the process: ϕ can be interpreted as a *roughness/complexity function* and characterizes the family to which the process belongs. Each class of functions ϕ defines a different kind of process (see examples just above), and inside each class the corresponding parameters (d, α, β, γ, $\ldots$) will be called the *complexity indexes*.

In light of what explained above, we propose a flexible approach to evaluate the complexity of functional data. The aim of our paper is twofold: firstly one has to detect the kind of process the data belong to (distinguishing between monomial and exponential families), and, once this is done, to estimate the complexity index(es). In the first step, starting from an estimate of ϕ, we introduce a method being *free of dominating measure* and based on a new graphical tool, named *log-Volugram*, that allows us to identify to which family of processes the statistical sample belongs (this is done along Section 2.1). To ensure a high degree of flexibility of the procedure, one has to use estimates being *free from parametric restriction* and models being *distribution-free*: to achieve this goal our procedure is based on kNN nonparametric functional smoothers which combine flexibility, easiness of implementation (because the dependence on a single discrete parameter) and location-adaptive feature. This is why the exploited kNN methodologies in functional data analysis are shortly reviewed at the beginning of Section 2. In the second step, once the class of the process is detected, the complexity index(es) (d, α, β, γ, $\ldots$) is(are) estimated and this can be done because of the free-modelling feature of the estimate ϕ. To do this we adopt a strategy commonly used in nonparametric framework: to study some specific submodel one compares a free-model estimate with what would be the true target under the submodel (see Härdle and Mammen, 1993 for earlier works in this direction in the multivariate regression setting). In our setting, the non-parametric estimates of ϕ is compared through a dissimilarity measure with one parametric family among the ones illustrated above and, by minimizing arguments, the complexity index(es) are estimated. This second step is presented in Section 2.2. Practical aspects about the introduced methodology and computational issues are discussed in Section 3.1 whereas the behaviour of the whole procedure is illustrated by means of a wide scope simulation studies in Section 3.2; these show good performances under different experimental conditions. Finally, to show how our two steps procedure can be usefully applied in a real case, we examine its performance in a financial framework to verify the compatibility of the data with standard model assumptions (see Section 4).

2. Methodology

In this section, after reviewing how the volumetric term in factorization (1) can be estimated nonparametrically, we show how to use it in developing new graphical tools that allow us to qualitatively detect the class of process from which the sample is drawn (see Section 2.1). Therefore, Section 2.2 describes how the nonparametric feature of the method allows to get estimates of the index complexity of the sample given the specified family.

The first statistical step consists in estimating both components in the decomposition (1) from a sample. To ensure a wide applicability of the method one has to develop statistical models/procedures being fully nonparametric. In the functional data setting, nonparametric statistics have been popularized in the book Ferraty and Vieu (2006) and are now widely used as long as one is interested in estimating some functional operator (regression, conditional distribution, ...). Among the various nonparametric smoothers, the kNN method is particularly adapted to the functional setting because it provides directly location adaptive estimates without needing highly complicated procedure (see Laloë, 2008; Burba, Ferraty and Vieu, 2009 for introductory works on functional kNN, see Biau, Cérou and Guyader, 2010; Lian, 2011; Kara et al., 2017; Kudraszow and Vieu, 2013 for the most recent advances and see Biau and Devroye, 2015 for a recent general presentation of kNN ideas).

Concerning the estimation of the terms in (1) the kNN estimates has a very simple and appealing form (see Ferraty et al., 2012). In fact, given a sample of n curves $X_1,\dots,X_n$ drawn from X, a point $\chi \in \mathscr{F}$ and a integer $k<n$, the surrogate density ψ at χ can be estimated by

$$\widehat{\psi}_k(\chi) = \frac{k(n-1)}{\sum_{i=1}^{n} k_i}, \tag{2}$$

where $k_i = \#\{j \neq i : X_j \in B(X_i, H_{n,k}(\chi))\}$, $H_{n,k}(\chi) = \min\left\{h \in \mathbb{R}^+, \sum_{i=1}^{n} 1_{B(\chi,h)}(X_i) = k\right\}$ and $1_A(x)$ is the characteristic function of the set A. As a matter of consequence, the single parameter involved in the method is a simple integer one, namely the number k of data contained in each neighbourhood.

At this stage, once the surrogate density is estimated and given the asymptotic factorization (1), one can easily derive nonparametric estimates of the volumetric component ϕ in the following way:

$$\phi_{k,n}(h) = \frac{n^{-1}\sum_{i=1}^{n} 1_{B(\chi,h)}(X_i)}{\widehat{\psi}_k(\chi)}. \tag{3}$$

Theoretical assessments related to the consistency of estimators (2) and (3) are developed in Ferraty et al., 2012. In order to compute (3) one has to face some practical problems. Firstly, since the asymptotic factorization (1) holds for small h, too large values must be avoided since they may increase the estimation error. At the same time, even too small values of h must be discharged since they force $\phi_{k,n}$ to be null: the ball, at the numerator on the right-side hand of (3), does not contain sample points. In other words, a suitable range of values $\mathscr{H} = [h_m, h_M]$ for h should be identified; for details see Section 3.1. Secondly, once h is appropriately chosen, one must take into account that the point χ, at which the SmBP is estimated, affects the approximation error of the whole factorization and, hence, the error of both $\widehat{\psi}_k$ and $\phi_{k,n}$. In this view, to circumvent such issue and to avoid an arbitrarily choice of χ, the estimation is averaged over the sample, that is

$$\widehat{\phi}_{k,n}(h) = n^{-1} \sum_{j=1}^{n} \phi_{k,n}^{(j)}(h), \tag{4}$$

where $\phi_{k,n}^{(j)}$ is (3) computed with $\chi = X_j$. In the following, if no ambiguities arise the dependences on k and/or n are dropped.

From such an estimate, one can visualize two graphical tools, that we name *Volugram* and *log-Volugram*. The shape of the latter is of help in discriminating among different family models for $\phi(\cdot)$ and in evaluating the *roughness/complexity indexes*. This is the basis of the descriptive approach to be developed in this paper.

2.1. The (log-)Volugram

The *Volugram* is the plot of $\widehat{\phi}$ computed on the realizations $x_1, \ldots, x_n$ versus h taken in a suitable positive interval sufficiently closed to zero. Because the quantities $\widehat{\phi}(h)$ are fully free from any kind of hypothesis (neither on the model, nor on the distribution of X, nor on any underlying dominating analytic structure), the observation of the shape of the curve $\widehat{\phi}$ can be directly used to have an idea on what is the complexity of the statistical sample. To fix the ideas let us just look at how behaves this Volugram in some simple examples. Figure 2 depicts the Volugrams of the noised 3-dimensional process and of the Brownian Bridge whose trajectories are illustrated in Figure 1. In both cases, estimations are based on samples of size $n = 200$, the number of neighbourhood is fixed to $k = n/2$ and, for the sake of computational practicality, h takes values in $\{H_{n,k}(x_i)\}_{i=1}^{n}$. Moreover, to ensure that the Volugram explores the smallest values of h, the plot is restricted to the 50% smallest values of the latter grid.

As is clear from Figure 2, although the Volugrams behave as one can expect (in both cases $\widehat{\phi}(h)$ decreases to zero as smaller values of h are considered), by looking at the sole Volugram it is not possible to discriminate from which family (exponential or monomial) the sample is drawn from. A practical tool to establish by eye such feature is instead provided by the *log-Volugram* defined as the plot of $\log \widehat{\phi}(h)$ versus $\log h$. Indeed, from a theoretical point of view, the volumetric term of processes in the monomial family satisfy, for small values of h, $\log \phi(h) \sim \alpha \log h$ whereas, in the exponential case, $\log \phi(h) \sim -C_2/h^{\beta}$. In other words, for small values of h, $\log \phi(h)$ is proportional to $\log h$ ($1/h^{\beta}$ respectively) for a process in the monomial (exponential respectively) family and the log-Volugram presents (does not present) a straight line shape. As a matter of

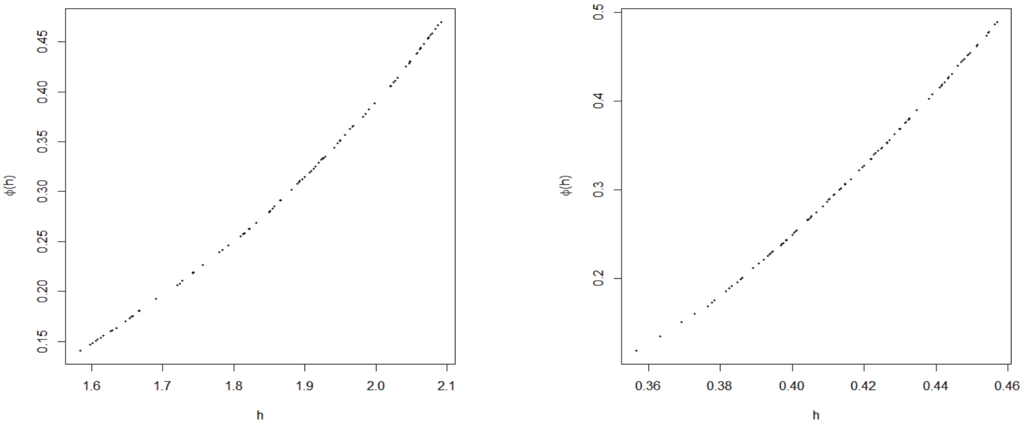

Figure 2: *Volugrams associated to a sample (of size 200) of a noised 3-dimensional process (left) and a Brownian Bridge (right) defined on* $[0,1]$*. In both cases,* $k = \lfloor n/2 \rfloor$ *and h takes values in the* 50% *smallest values of* $\{H_{n,k}(x_i)\}_{i=1}^{n}$.

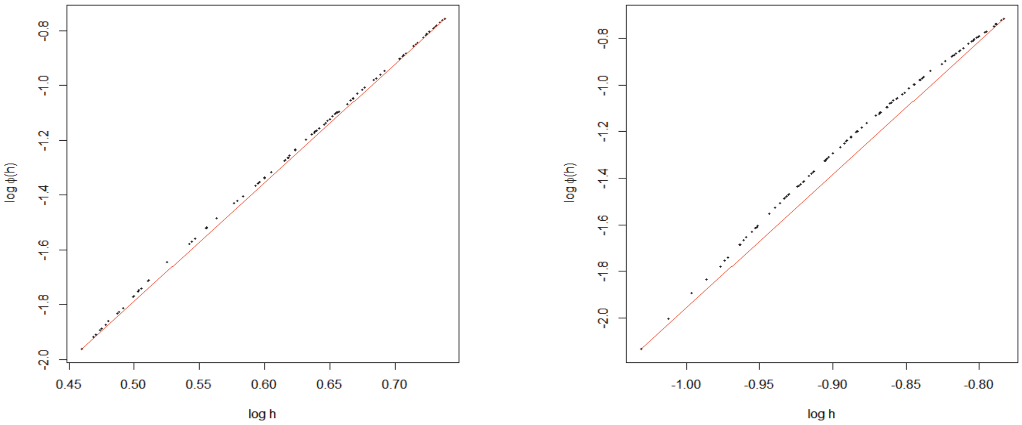

Figure 3: *log-Volugrams associated to a sample (of size 200) of a noised 3-dimensional process (left) and a Brownian Bridge (right) defined on* $[0,1]$*. In both cases,* $k = \lfloor n/2 \rfloor$ *and h takes values in the* 50% *smallest values of* $\{H_{n,k}(x_i)\}_{i=1}^{n}$*. The line passing through the first and the last points (ordered according to the ascending order of h) is drawn as well.*

illustration and using the same data and settings of Figure 2, the correspondent log-Volugrams are depicted in Figure 3. For the sake of comparison, the latter figures are completed by overlapping the line passing through the first and the last points (ordered according to the ascending order of h).

These arguments make clear how the log-Volugram allows, better than the Volugram, to drive the researcher towards the family of processes from which the sample comes. In particular, the more $\{(\log h, \log \widehat{\phi}(h))\}$ are aligned, the greater the compatibility to the monomial model is. On the contrary, deviations from this situation represent an empirical evidence of exponential model. Hence, one can decide that the theoretical volumetric function $\phi(\cdot)$ is of some specific form depending on a complexity parameter $\theta \in \Theta$ where Θ is a subset of $\mathbb{R}^p$

$$\phi \in \mathscr{C} = \{\phi_\theta, \theta \in \Theta\}.$$

To fix the idea, the left panel of Figure 3 suggests the monomial family $\mathscr{C}_M = \{\phi_\alpha(h) = c_\alpha h^\alpha, \alpha > 0\}$, whilst the right panel leads towards the exponential one $\mathscr{C}_E = \{\phi_{(\gamma,\beta)}(h) = C_1 h^\gamma \exp\{-C_2/h^\beta\}, \beta > 0, \gamma \geq 0\}$.

2.2. Estimating the complexity index

In the second step of the procedure the aim is to gain more insights into the structure of the data by intending to estimate the complexity index θ of the chosen family $\mathscr{C}$ by means of a comparison between the free-model estimate ϕ with one of the parametric family that would be the true target. Precisely, this leads to consider the *centered cosine dissimilarity* between $g(\phi_\theta)$ and $g(\widehat{\phi}_k)$ computed on the observed values and defined by

$$\Delta(\widehat{\phi}_k, \phi_\theta) = 1 - \frac{\langle \widetilde{g}(\phi_\theta), \widetilde{g}(\widehat{\phi}_k)\rangle^2}{\|\widetilde{g}(\phi_\theta)\|^2 \|\widetilde{g}(\widehat{\phi}_k)\|^2}, \qquad k = 1, 2, \ldots, (n-1),\ \theta \in \Theta, \tag{5}$$

where $\langle f_1, f_2 \rangle = \int_{\mathscr{H}} f_1 f_2$ with $\mathscr{H}$ being a suitable interval included in $(0, \infty)$, $\|f\|^2 = \langle f, f \rangle$ and $\widetilde{g}(\phi) = g(\phi) - \int_{\mathscr{H}} g(\phi)$ with $g(\cdot)$ a suitable continuous real valued function defined on $(0, +\infty)$. Note that centered cosine dissimilarity is invariant for affine transformations. Practical aspects in computing (5), including how $g(\cdot)$ and $\mathscr{H}$ are chosen, are treated in details in Section 3.1. The idea is to estimate the complexity index that minimizes $\Delta(\widehat{\phi}_k, \phi_\theta)$ over suitable grids $\mathscr{T}$ for θ, and $\mathscr{K}$ for k. Let us now show how such dissimilarity behaves in the simple examples that are following through the paper. Figure 4 depicts the heat-map of Δ (top panels) and the curves $\{\Delta(\widehat{\phi}_k, \phi_\theta) : k \in \mathscr{K}\}$ (bottom panels). These heuristically show that Δ reaches a minimum which appears rather stable with respect to the choice of k.

That spontaneously leads to estimate the complexity index by minimizing (5) for a fixed k, that is

$$\widehat{\theta} = \arg\min_{\theta \in \mathscr{T}} \Delta(\widehat{\phi}_k, \phi_\theta).$$

At this stage, it is worth noticing that if the shape of log-Volugram produces doubts in the choice of the family, it is always convenient to firstly classify the sample as drawn from the exponential family and estimate β. If a misspecification of the model occurred, then the estimation of the complexity index tends to assume the minimum values in the grid $\mathscr{T}$, see Figure 5; i.e. the exponential part of the volumetric term can be considered negligible. This can be used as a feedback procedure to avoid this kind of misspecification error.

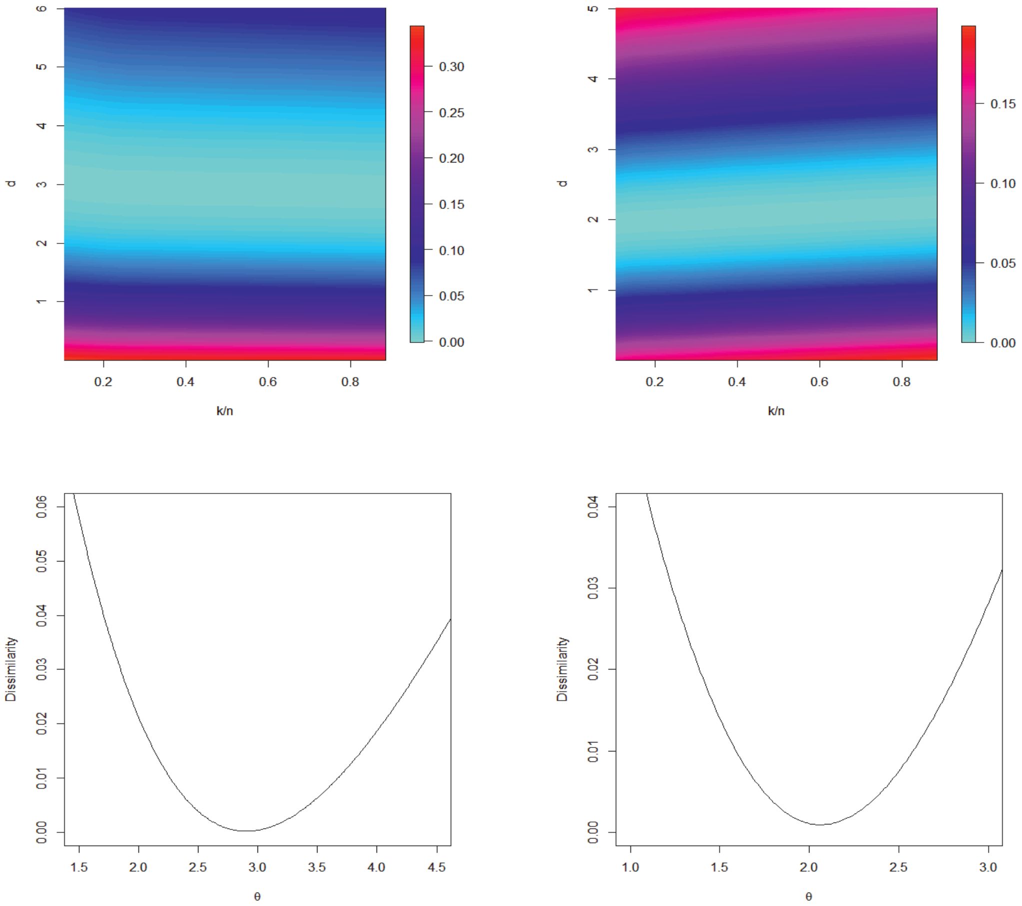

Figure 4: *Top panels: the heat maps of* $\Delta(\widehat{\phi}_k, \phi_\theta)$ *as a function of* k/n *and* θ *associated to a sample (of size 200) of a noised 3-dimensional process (left) and a Brownian Bridge (right) defined on* $[0,1]$. *Bottom panels: graphs of* $\Delta(\widehat{\phi}_k, \phi_\theta)$ *with* $k = \lfloor n/2 \rfloor$, *as a function of* θ, *associated to the same samples.*

In conclusion, the method detects the good class as explained above and, within the selected family, it seems also capable to find a good estimation of the complexity index. Simulations described in what follows confirm these abilities.

3. Algorithm in action

In this section we firstly describe the algorithm in Section 3.1 and, soon after in Section 3.2, we show its performance over a set of selected simulations under different experimental conditions.

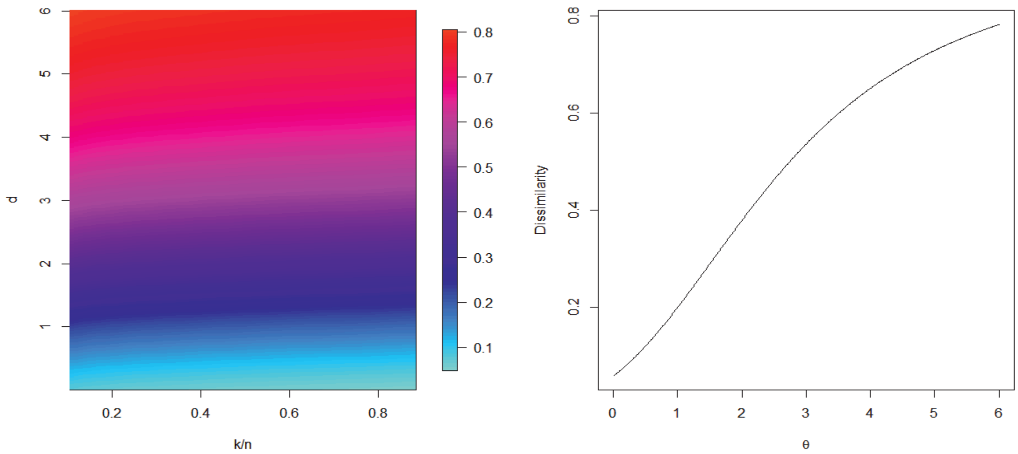

Figure 5: *The heat map of $\Delta(\widehat{\phi}_k, \phi_\theta)$ (left panel) and the plot of $\Delta(\widehat{\phi}_{k_0}, \phi_\theta)$ against θ (right panel) when the noised 3-dimensional process (in left panel of Figure 3) is confused with a process from the exponential family.*

3.1. Procedure description

Here we detail the algorithm features; some of them depend on the family identified at the first step of the methodology as described in Section 2.1.

Although the algorithm could be implemented for potentially any $\mathscr{F}$ whose topology is induced by a semimetric ρ, here, for simplicity, $\mathscr{F}$ is $\mathscr{L}^2_{[0,1]}$: the separable Hilbert space of square integrable function on $[0,1]$ with usual inner product, norm and induced metric. Thus the realizations $x_1, \ldots, x_n$ of a sample $X_1, \ldots, X_n$, drawn from the $\mathscr{F}$-valued random element X, are considered.

In computing the dissimilarity measure $\Delta(\widehat{\phi}_k, \phi_\theta)$, we have to specify $g(\cdot)$, $\mathscr{H}$, $\mathscr{T}$ and $\mathscr{K}$.

For what concerns the transformation $g(\cdot)$, if the monomial class $\mathscr{C}_M$ is suggested by the log-Volugram, g is the identity function, whereas, for the exponential class $\mathscr{C}_E$, it is the logarithm transformation. In both cases, the transformed empirical volumetric term $g(\widehat{\phi})$ is then compared with a term in the simple form ch^θ for small values of h. In fact, if $\phi \in \mathscr{C}_M$, then $\phi(h) = c_\alpha h^\alpha$ with $\alpha \in (0, \infty)$. If $\phi \in \mathscr{C}_E$,

$$\log \phi(h) = \log C_1 + \gamma \log h - C_2 h^{-\beta} \sim -C_2 h^{-\beta} \tag{6}$$

and then, in the exponential case, the leading complexity parameter is β. Indeed, at the best of our knowledge, for the most of processes related to Brownian motion with known SmBP asymptotic, it holds $\log \phi(h) \sim -C_2 h^{-2}$ (see, for instance, Nikitin and Pusev, 2013). In particular, $C_2 = 1/8$ when X is Wiener, Brownian Bridge (BB), Geometric

Brownian Motion (GBM), Ornstein-Uhlenbeck. Anyway, note that (6) is more accurate if $\gamma = 0$, and this happens, for instance, in the case of Brownian Bridge that consequently becomes a benchmark process. In practice, beside the BB, we have specialized our method to deal with those processes suspected to be Wiener or GBM since these can be led back to a BB by means of suitable transformations. In more details, if $X(t) = W(t)$ is Wiener on $t \in [0,1]$, then

$$W(t) - tW(1), \tag{7}$$

is a BB on $[0,1]$, whereas if $X(t)$ is the GBM identified by the stochastic differential equation

$$\begin{cases} dX(t) = \mu X(t)dt + \sigma X(t)dW(t), & t \in [0,1], \\ X(0), \sigma > 0, \end{cases} \tag{8}$$

whose solution is $X(t) = X(0)\exp\left\{\left(\mu - \sigma^2/2\right)t + \sigma W(t)\right\}$, $t \in [0,1]$, then

$$[\log(X(t)/X(0)) - (\mu - \sigma^2/2)t]/\sigma, \qquad t \in [0,1] \tag{9}$$

is a Wiener process for which transformation (7) can be applied, leading to a BB on $[0,1]$. The estimation of γ remains an open problem for processes different from the BB, the Wiener process and the GBM.

For what concerns $\mathscr{H} = [h_m, h_M]$, h_m is chosen in order to guarantee that there exists at least an observed curve x_i for which $B(x_i, h_m)$ includes some $x_j \neq x_i$; whereas, the range $\mathscr{H}$ should become closer to zero as the sample size increases.

Finally, $\mathscr{T}$ is a equally spaced mesh over an interval that varies with the experimental setting; our suggestion is to start with a wide range of values with a relatively rough step, then to restrict the region of search by using a finer grid. To reveal possible dependencies on k, in the simulation study, we use $\mathscr{K} = \{\lfloor \delta n \rfloor : \delta = 1/4, 1/3, 1/2\}$ with $\lfloor \delta n \rfloor$ being the smaller integer greater than δn. Such a choice is coherent with many rules introduced in literature (see, for instance, Devroye, Györfi and Lugosi, 1996; Duda, Hart and Stork, 2012; Györfi et al., 2006).

3.2. Numerical Experiments

In this section we present the results of numerical experiments aimed to evaluate the ability of the method in estimating the complexity parameter by varying the underlying process and the sample size.

We generate 1000 Monte Carlo samples each one constituted by n independent random curves $X_1, \ldots, X_n$ drawn from a process X with $n = 50, 100, 200, 500$. From each sample the complexity index is estimated and its distribution analysed. In particular, we consider noised finite dimensional processes and infinite dimensional ones.

About the noised finite dimensional processes, curves are generated according to

$$X(t) = \sum_{j=1}^{d} a_j \xi_j(t) + \mathscr{E}(t), \qquad t \in [0,1]$$

where $\{\xi_j\}_{j=1}^{d}$ are the first d elements of the Fourier basis

$$\xi_j(t) = \begin{cases} \sqrt{2}\sin(2\pi m t - \pi), \ j = 2m-1 \\ \sqrt{2}\cos(2\pi m t - \pi), \ j = 2m \end{cases} \qquad m \in \mathbb{N},$$

$\{a_j\}_{j=1}^{d}$ are i.i.d. as $\mathscr{N}(0,1)$ and $\mathscr{E}(t)$ is a Gaussian white noise with $\sigma = 0.02$ representing a measurement error. Here, $d = 3$ and $d = 6$ are considered.

For what concerns the infinite dimensional processes, we consider the Wiener process and the Geometric Brownian Motion (GBM). Each curve in both families are discretized over an equispaced grid on $[0,1]$ consisting of 100 points: the resulting data-sets are $n \times 100$ matrices with entries $x_{i,j}, i = 1,\dots,n, j = 1,\dots,100$. GBM trajectories are simulated from the stochastic differential equation (8), with $X(0) = 1$, $\mu = 0$, $\sigma = 1$, using the Euler-Maruyama approximation scheme (Kloeden and Platen, 1992, Section 9.1). Coherently with what stated in the previous section, Wiener curves are transformed by means of (7) whereas GBM trajectories by (9) and, successively, (7). To operationalize (9), maximum likelihood estimates of parameters are computed for each discretized curve in the sample: for each $i = 1,\dots,n$, μ and σ are estimated by $\hat{\mu} = 100^{-1}\sum_{j=1}^{100} x_{i,j}$ and $\hat{\sigma}^2 = 100^{-1}\sum_{j=1}^{100}(x_{i,j} - \hat{\mu})^2$ respectively. Integrals in (5) are approximated on such grid by using a rectangular numerical rule. In what follows, $\mathscr{T}$ is an equispaced grid with step 0.01 and $k = \lfloor \delta n \rfloor$ with $\delta = 1/4, 1/3, 1/2$.

Table 1 collects the results from the Monte Carlo experiments from which we can appreciate the good performances of complexity index estimator. In particular, in all the cases no relevant bias arises, variability of the estimator is moderate, especially, in relative terms with respect to the true parameter. As expected, variability decreases with n whereas, in the finite dimensional case, it slightly increases with the complexity: the larger d is, the larger the variability in relative terms with respect to the true parameter is. These comments hold true for all the chosen k, therefore, for practical purposes, an heuristic choice like $k = \lfloor n/2 \rfloor$ is reasonable.

The distributions of estimated values $\widehat{d}$ and $\widehat{\beta}$ over the 1000 simulations when $n = 500$ and $k = 250$ are plotted in Figure 6: dashed vertical lines are superimposed to kernel density estimates in correspondence of extreme quantiles of order 0.025 and 0.975, in order to delimit a Monte Carlo empirical 95% confidence interval. All distributions appear rather symmetric and bell-shaped: anyway, the Shapiro-Wilk test tends to reject the normality assumption in all the cases at the level 5%.

Table 1: *Synthetic indicators of the estimated complexity indexes obtained from 1000 MC replication under different experimental conditions.*

Process	$\delta \rightarrow$	1/2		1/3		1/4	
family	$n \downarrow$	Mean	St.dev	Mean	St.dev	Mean	St.dev
Finite dimensional	50	2.966	0.305	2.994	0.312	3.023	0.316
(with $d = 3$)	100	2.985	0.167	3.007	0.169	3.023	0.170
	200	3.014	0.100	3.034	0.100	3.046	0.100
	500	3.072	0.072	3.090	0.071	3.101	0.070
Finite dimensional	50	5.938	1.559	5.962	1.596	5.986	1.629
(with $d = 6$)	100	5.891	0.911	5.926	0.933	5.947	0.947
	200	5.894	0.497	5.923	0.509	5.940	0.521
	500	5.985	0.249	6.008	0.255	6.023	0.260
GBM ($\beta = 2$)	50	1.995	0.190	1.927	0.187	1.897	0.188
	100	2.011	0.117	1.949	0.113	1.916	0.112
	200	2.015	0.072	1.956	0.070	1.924	0.069
	500	2.005	0.039	1.952	0.038	1.921	0.038
Wiener ($\beta = 2$)	50	1.987	0.190	1.920	0.187	1.888	0.188
	100	2.005	0.117	1.942	0.113	1.909	0.112
	200	2.009	0.071	1.950	0.069	1.918	0.069
	500	1.999	0.039	1.946	0.038	1.915	0.037

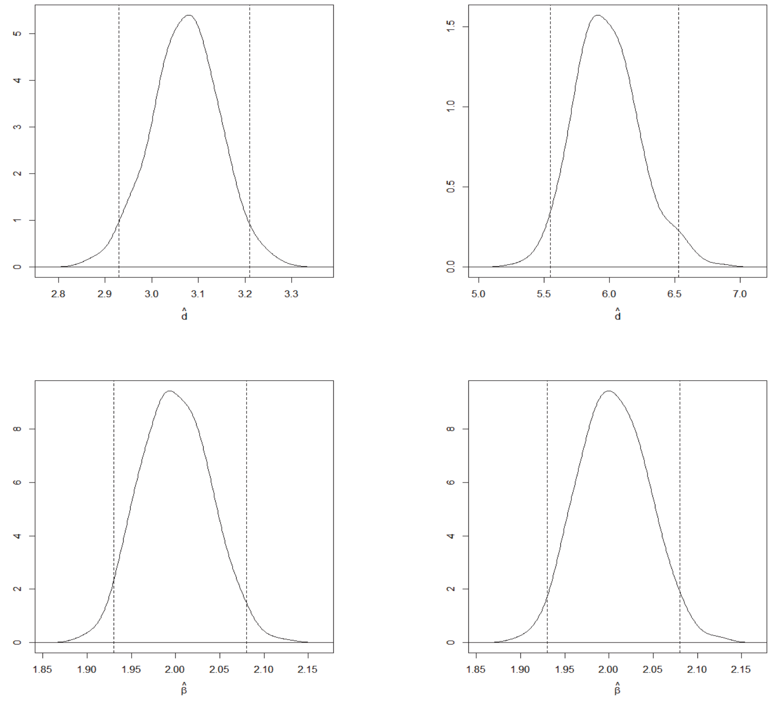

Figure 6: *Kernel density estimates of $\widehat{d}$ and $\widehat{\beta}$ for the finite dimensional processes ($d = 3$ and $d = 6$, left and right top panels respectively) and for the GBM and Wiener processes ($\beta = 2$, left and right bottom panels respectively) when $n = 500$. Dashed vertical lines correspond to the 95% Monte Carlo confidence interval limits.*

4. Application to financial data

A common problem in finance is the modelling of stock prices time series, for example in implementing parametric option pricing models via Monte Carlo simulations. Thanks to its nice properties, the GBM has gained a central place in theoretical and applied financial literature, becoming a prototype for a generation of models; see, for example, Fusai and Roncoroni (2007) and Campbell, Lo and MacKinlay (1997).

In this section we illustrate how the proposed methodology can provide a tool for practitioners in detecting the family of processes to which the observed time series belongs, and for a rough evaluation of the complexity of such data. To do this, we analyze in details the case of the S&P500 during the period 14th October 2016, 15th January 2017 with 1 minute frequency for a total of 63 market days and 390 observations per day (we deleted shorter days). Data are collected by using the link https://www.google.com/finance/getprices?i=60&p=200d&f=d,o,h,l,c,v&df=cpct&q=.INX. The corresponding trajectory is depicted in the left panel of Figure 7. To qualitatively assess that the observed trajectory is compatible with a GBM process, we apply our method on a sample derived from above dataset: given the high frequency of measurements, each market day is divided into three non-overlapping parts having the same size to which correspond three trajectories. Consequently, the sample is formed by $n = 189$ each one discretized over an equally spaced grid of 130 points.

In order to implement the two steps of our method, the sample must be transformed as explained in detail in Section 3.2. In particular, given the assumption that the underlying process is a GBM, since drift and volatility of a stock process vary with time, it is reasonable to model each curve x_i with specific parameters μ_i and σ_i. They are estimated by using the maximum likelihood approach illustrated in the previous section,

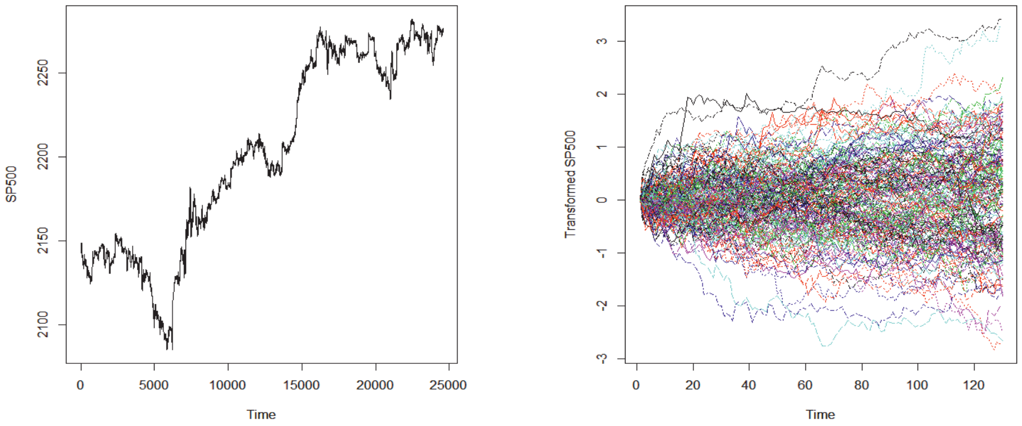

Figure 7: Left panel - Trajectory of S&P500 value from 14th October 2016 to 15th January 2017 with 1 minute frequency. Right panel - The functional sample: each functional observation is one third of a market day trajectory after transformation (9).

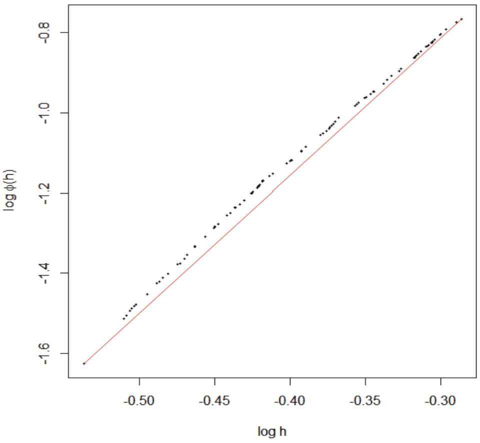

Figure 8: *The log-Volugram for the transformed S&P500 sample.*

starting from discretized points of each curve. The sample of curves which arise from these manipulations is plotted in the right panel of Figure 7.

In the same spirit of Section 2.1, we plot the log-Volugram with $k = \lfloor n/2 \rfloor$, see Figure 8. Its shape drives our analysis towards the exponential family. The heat map of dissimilarity Δ and the dissimilarity computed at $k = \lfloor n/2 \rfloor$ are drawn in Figure 9. The minimization of Δ leads to $\widehat{\beta} = 1.94$.

This first analysis supports the assumption that S&P500 could be modelled as a GBM with varying parameters at least for a short time period.

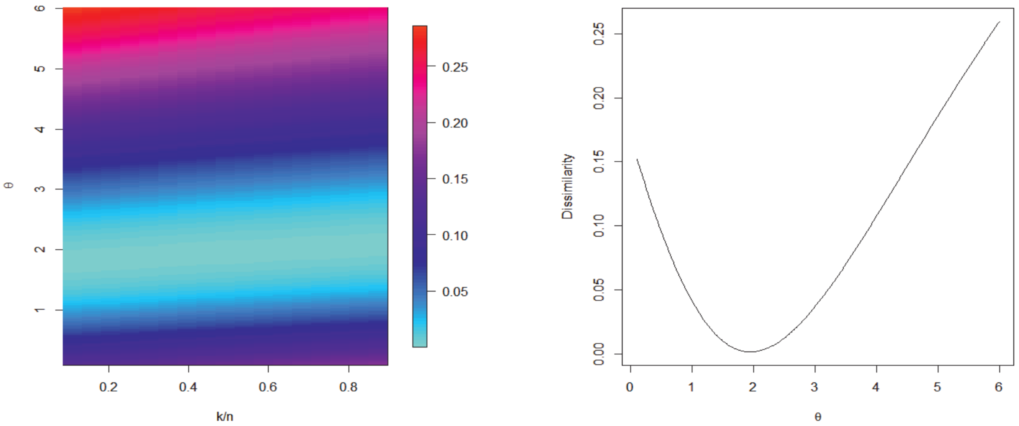

Figure 9: *The heat map of Δ (left panel) and Δ at $k = \lfloor n/2 \rfloor$ (right panel) for the transformed S&P500 sample.*

In order to evaluate the stability of results with respect to the way in which we built the sample of functional data, we repeated the analysis using different cutting criteria: besides dividing each market day in three parts, we tried also with two parts consisting of 195 points, five parts of 78 points and 6 parts of 65 (all the intervals are not overlapped). Resulting samples have sizes $n = 126, 315, 378$ whereas the results obtained with $k = \lfloor n/2 \rfloor$ are $\widehat{\beta} = 1.94, 1.96, 1.98$ respectively. They confirm the compatibility of data with a GBM (with time varying parameters) assumption that, hence, can be used as a good approximating model for performing option pricing.

5. Comments

This paper has provided flexible tools for analysing the complexity of a functional statistical sample. In order to ensure its high degree of applicability the procedure is free from any structural assumption from several points of view: from an analytic point of view (it is free from any dominating measure assumption in the underlying infinite dimensional space), from a probabilistic point of view (it is free from any distribution assumption on the underlying stochastic process), from a statistical point of view (it is free from any parametric assumption on the model), and from a computational point of view (the method depends on a single discrete parameter). This has been possible by using kNN ideas that combine good theoretical properties and ease of implementation. In a first step, the method provides some graphical tools (the so-called Volugram or log-Volugram) which are used to detect the class of complexity of the data, while in a second step it provides an automatic estimate of the index of complexity inside of the detected class. The methodology provides excellent results in evaluating the complexity family and index on simulated and real datasets.

Acknowledgements

We wish to thank the Associate Editor and an anonymous referee for their valuable comments. This work was partially developed during a visiting research period at the Institut de Mathématiques de Toulouse, Université Paul Sabatier. E. Bongiorno and A. Goia thank for the hospitality. E. Bongiorno and A. Goia are members of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM). The financial support of CRoNoS - COST Action IC1408 is acknowledged.

References

Aneiros, G., Bongiorno, E.G., Cao, R. and Vieu, P. (2017). *Functional Statistics and Related Fields*. Springer.

Biau, G., Cérou, F. and Guyader, A. (2010). Rates of convergence of the functional k-nearest neighbor estimate. *IEEE Transactions on Information Theory*, 56, 2034–2040.

Biau, G. and Devroye, L. (2015). *Lectures on the Nearest Neighbor Method.* Springer Series in the Data Sciences. Springer, Cham.

Bogachev, V.I. (1998). *Gaussian Measures.* Vol. 62 of Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI.

Bongiorno, E.G. and Goia, A. (2016). Classification methods for Hilbert data based on surrogate density. *Computational Statistics & Data Analysis*, 99, 204–222.

Bongiorno, E.G. and Goia, A. (2017). Some insights about the small ball probability factorization for Hilbert random elements. *Statistica Sinica*, Forthcoming.

Bongiorno, E.G., Goia, A., Salinelli, E. and Vieu, P. (Eds.) (2014). *Contributions in Infinite-Dimensional Statistics and Related Topics.* Società Editrice Esculapio.

Bosq, D. (2000). *Linear Processes in Function Spaces.* Vol. 149 of Lecture Notes in Statistics. Springer-Verlag, New York.

Burba, F., Ferraty, F. and Vieu, P. (2009). k-nearest neighbour method in functional nonparametric regression. *Journal of Nonparametric Statistics*, 21, 453–469.

Campbell, J.Y., Lo, A.W.-C. and MacKinlay, A.C. (1997). *The Econometrics of Financial Markets.* Princeton University Press.

Cardot, H., Cénac, P. and Godichon-Baggioni, A. (2017). Online estimation of the geometric median in Hilbert spaces: Nonasymptotic confidence balls. *Annals of Statistics*, 45, 591–614.

Chen, K., Delicado, P. and Müller, H.-G. (2017). Modelling function-valued stochastic processes, with applications to fertility dynamics. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 79, 177–196.

Ciollaro, M., Genovese, C., Lei, J. and Wasserman, L. (2014). *The Functional Mean-Shift Algorithm for Mode Hunting and Clustering in Infinite Dimensions*. Preprint.

Delaigle, A. and Hall, P. (2010). Defining probability density for a distribution of random functions. *Annals of Statistics*, 38, 1171–1193.

Delsol, L. and Louchet, C. (2014). Segmentation of hyperspectral images from functional kernel density estimation. In: *Contributions in Infinite-Dimensional Statistics and Related Topics.* Esculapio, Bologna, pp. 101–106.

Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Vol. 31 of Applications of Mathematics (New York). Springer-Verlag, New York.

Duda, R.O., Hart, P.E. and Stork, D.G. (2012). *Pattern Classification.* John Wiley & Sons.

Ferraty, F., Kudraszow, N. and Vieu, P. (2012). Nonparametric estimation of a surrogate density function in infinite-dimensional spaces. *Journal of Nonparametric Statistics*, 24, 447–464.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis.* Springer Series in Statistics. Springer, New York.

Fusai, G. and Roncoroni, A. (2007). *Implementing Models in Quantitative Finance: Methods and Cases.* Springer Science & Business Media.

Gasser, T., Hall, P. and Presnell, B. (1998). Nonparametric estimation of the mode of a dis-tribution of random curves. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 60, 681–691.

Goia, A. and Vieu, P. (2016). An introduction to recent advances in high/infinite dimensional statistics [Editorial]. *Journal of Multivariate Analysis*, 146, 1–6.

Györfi, L., Kohler, M., Krzyzak, A.and Walk, H. (2006). *A Distribution-Free Theory of Non-Parametric Regression*. Springer Science & Business Media.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Annals of Statistics*, 21, 1926–1947.

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*. Springer Series in Statistics. Springer, New York.

Jacques, J. and Preda, C. (2014). Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8, 231–255.

Kara, L.-Z., Laksaci, A., Rachdi, M. and Vieu, P. (2017). Data-driven *k*NN estimation in nonparametric functional data analysis. *Journal of Multivariate Analysis*, 153, 176–188.

Kloeden, P.E. and Platen, E. (1992). *Numerical Solution of Stochastic Differential Equations*. Vol. 23 of Applications of Mathematics (New York). Springer-Verlag, Berlin.

Kokoszka, P., Oja, H., Park, B. and Sangalli, L. (2017). Special issue on functional data analysis. *Econometrics and Statistics*, 1, 99–100.

Kudraszow, N.L. and Vieu, P. (2013). Uniform consistency of *k*NN regressors for functional variables. *Statistics & Probability Letters*, 83, 1863–1870.

Laloë, T. (2008). A *k*-nearest neighbor approach for functional regression. *Statistics & Probability Letters*, 78, 1189–1193.

Li, W.V., Shao and Q.-M. (2001). Gaussian processes: inequalities, small ball probabilities and applications. In: *Stochastic Processes: Theory and Methods*. Vol. 19 of Handbook of Statistics North-Holland, Amsterdam, pp. 533–597.

Lian, H. (2011). Convergence of functional *k*-nearest neighbor regression estimate with functional responses. *Electronic Journal of Statistics*, 5, 31–40.

Lifshits, M.A. (2012). *Lectures on Gaussian Processes*. Springer Briefs in Mathematics. Springer, Heidelberg.

Masry, E. (2005). Nonparametric regression estimation for dependent functional data: asymptotic normality. *Stochastic Processes and their Applications*, 115, 155–177.

Nikitin, Y.Y. and Pusev, R.S. (2013). Exact small deviation asymptotics for some Brownian functionals. *Theory of Probability and Its Applications*, 57, 60–81.

Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*, 2nd Edition. Springer Series in Statistics. Springer, New York.

Vilar, J.M., Raña, P. and Aneiros, G. (2016). Using robust FPCA to identify outliers in functional time series, with applications to the electricity market. *SORT*, 40, 321–348.